DOCUMENT RESUME

ED 266 692 FL 015 548

AUTHOR Stevenson, Douglas K.
TITLE The Teaching of Testing: Reliability, Validity, and All That Stuff.
PUB DATE 85
NOTE 16p.; In: Language Testing in School. AFinLA Yearbook 1985. No. 41 (see FL 015 537). Document is in small print.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Foreign Countries; Inservice Teacher Education; Instructional Improvement; Language Research; *Language Teachers; *Language Tests; Second Language Instruction; Teacher Education; Teaching Methods; *Testing; Test Reliability; Test Validity

ABSTRACT

While language testing specialists agree on the importance of their field and that the intelligent application of language testing approaches, techniques, and methods to the classroom can improve the overall quality of language learning and teaching, that opinion may not be shared by teachers. Few teachers have training in language testing, and while there is appropriate literature available to teachers, they must use it to gain from it. The trend toward accountability and the growing tendency to doubt the teacher's word argue in favor of better testing. Four areas of testing deserve increased emphasis in teacher training: the concepts of reliability and validity; the overall educational and social context; language learning research; and ethics, accountability, and fairness. While reliability and validity have a reputation for being esoteric, an understanding among teachers of these concepts and of their interrelationship clarifies the importance of testing and can do more than anything else for its improvement. Awareness and understanding of each of the four areas are highly useful in the classroom, and the quality of testing stands to gain as much from the teaching of testing as from more testing. (MSE)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Tommola

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED266692

# THE TEACHING OF TESTING: RELIABILITY, VALIDITY, AND ALL THAT STUFF!

Douglas K. Stevenson (University of Essen)

Two years ago - almost exactly to the day - I was in New York City at an ACTFL conference (that is, the American Council on the Teaching of Foreign Languages). Some friends and I were standing in a hallway outside a room, nervously awaiting our turn to speak. We'd been asked to be members of a panel, and the topic for our panel was practical applications of research in foreign language teaching. I'd been asked to take the language testing part; in other words, to deal with the question, "what does language testing research have to offer the foreign language teacher?" As I've said, we were a bit nervous. It was, after all, a very good question, and we were worried that someone in the audience might actually ask it.

While we were standing there, trying to think of some appropriate answers, two teachers came by. They were busily discussing that morning's program: which of several lectures they wanted to attend, which of the many panel discussions sounded interesting, or which demonstrations of new teaching techniques might be worth seeing. Needless to say, we were all listening in on their conversation, eavesdropping to hear their opinions. Then, after some discussion, one of the teachers pointed to a lecture in the program on a language testing topic. I've forgotten the exact title of the lecture, but it was something like "designing oral achievement tests for the classroom." "Now, that might be interesting," the teacher said. "Oh, no, we don't want to hear that, do we?" the other one replied. "They'll probably just be talking about reliability, validity and, you know, all that stuff!" Well, we thought, and laughed, that pretty much takes care of language testing, at least for them.

I'd like to talk to you this morning about an important area, an area that doesn't usually get the attention it deserves. This is the teaching of language testing to teachers. I'd like to first mention some difficulties we have, to sketch out the present situation. Then, in the second half of the paper, I'd like to talk about some content areas that need to be stressed more than they have been. Now, while I've thought about this topic, that mini-dialogue between two teachers, overheard two years ago in New York, has constantly been in my mind. It is there, itching at my memory, whenever the subject of language testing at school comes up. For a very good reason, I think, a reason we language testers tend to forget or, perhaps, would like to. Let me explain it this way.

FL 015548

Some of us here this morning might be described as language testing specialists (in addition, that is, to being teachers). Not necessarily experts, mind you. Rather, individuals - whether working primarily in the schools or universities, with testing centers or governmental agencies - who spend a great portion of their professional lives being actively and immediately concerned with the field of language testing in all of its many aspects.

Those of you who fit this description have probably by now spent several years just reading the language testing literature, those many articles, studies, research reports and proposals, dissertations and theses, conference proceedings, journals and books that are all related, directly or indirectly, to language testing. And, as you are only too well aware, this already vast literature keeps growing month by month and year by year. And, it's not just in Europe or even in North America that so much appears about language testing and assessment. For example, there's been the notably well-attended conferences in Singapore and in Hong Kong, the latest in a series. Moreover, what is being done in Japan, or in the Chinas, is also of considerable interest, especially to those of us concerned with the teaching and testing of International English. But whatever the specific area within the broad field of language testing, and from wherever it comes, be it Essen or Essex, Taiwan or Tampere, it will necessarily be of some interest to language testing specialists. To language testing specialists, language testing is, but of course, the major interest.

On the other hand, language testing specialists - like all specialists - are often so enamored of their subject, so fascinated by its complexities - that they overlook a simple thing. They tend to forget that their primary field of interest is not the ultimate academic delight for everyone else. And, I'm sorry to say, even in linguistics and language pedagogy there are those who feel, and have sometimes even stated, that the language teaching world does not revolve around language testing theory and practice.

We know that it should, of course. And, that's not just a feeling that derives from academic ethnocentrism, or so-called academic imperialism. After all, as Upshur and Oller and many other testers have pointed out, the assessment, measurement, and testing of language and language-related behavior - however conceived and carried out - is at the very center of all linguistics, language teaching, and language learning research.

Nonetheless, the fact that language testers often state this does not mean that others see it that way or, for that matter, are even disposed to consider it. Rather, they too as specialists are inclined to see their fields, their specialties, as constituting what language teachers really need. In fact, just about everyone in linguistics and language pedagogy is fighting for the attention of the classroom language teacher. Phoneticians insist that our future teachers must know a great deal about the etics and emics of their field. Sociolinguists, too, argue for more time with future teachers: it is obvious, isn't it, that sociolinguistic appropriateness and sensitivity bear at least as much weight in communicative situations as do accent and pronunciation? Even our colleagues who specialize in cultural studies will state their case to anyone willing to listen, and especially to the authorities who have required teachers to take courses in literature, linguistics, and didactics. Without cultural knowledge, they argue, these three areas produce only shallow pools of perfomance.

In this crowded, loud bazaar of language-based disciplines, the language tester has a very small voice, indeed, when it comes to attracting and holding the attention of the language teacher. What it

comes down to, I guess we could say, is the somewhat dismal fact that no matter how fascinating, and multifaceted, and fundamentally important language testing is to the language tester, it simply does not have, and never has had, the same inherent interest or felt-importance to the rest of the language teaching fraternity. And so, we might add, our sympathies go to them.

Now I do not find this situation, that language testing is not everyone's cup of tea, especially shocking. Nor would any of us, looking back over the past twenty years or so since Lado's Language Testing appeared (1961), expect the situation to change radically. It is simply the standard state of affairs, most everywhere, more or less, and if many of us are not content to live with it, we at least admit that it is there and should be changed. However, as I have suggested thus far, the topic does invite attention. There are several very basic reasons for this, each related to our central goal in language testing, which is to improve the quality of language assessment overall. Let me present the three most important of these, which are best considered as functioning assumptions.

First, you here today, that is, those of you who would not consider yourselves language testing specialists, do form a fairly unusual group. You are unusual, in a most positive sense, I hasten to add, because you are obviously willing to spend not a little of your time and schedules discussing language testing at school. There have been of course other such meeting and seminars. The IUS has sponsored some in cooperation with other groups. Furthermore, organizations such as International TESOL, or ACTFL, the British Council, and CITO have encouraged such gatherings. Yet all in all, it is still far from common to find groups of teachers voluntarily coming together to talk about testing. It is also still true that only a few of the hundreds of thousands of practicing teachers of modern languages throughout the world have had a chance to take any specific coursework in language testing. Usually, what is taken into the classroom is at most a few lectures. As I've discussed this situation in detail elsewhere (1983), I won't repeat it here. It can be concluded, however, that overall, language testing continues to be treated as if it were a nice, but not really necessary, subject.

Here we should remember that today's teachers are expected to be at home in a very wide range of disciplines and somehow - no one seems to spend much time wondering exactly how - they are expected to acquire all this knowledge within the little time usually available. It is, of course, unreasonable to expect someone to know much about what they have not been taught, or had a chance to learn. Simply to state, as is so often done, that "all good teachers" should know such and such, be it language testing or any other area, is to confuse an ideal syllabus with actual achievement. It all reminds me of a particularly humorless history professor we endured as undergraduates. When someone with enough courage finally asked him what would be included on his final examination, a fearsome four-hour affair, he gave a dry smile and said, "anything I haven't covered in class."

The first point, therefore, is that few teachers will be taught much about language testing during their teacher training period, and yet they are still somehow expected to know more than a little about it as classroom teachers. And certainly, whether or not language testing is a required subject in teacher training, and most often it is not, all of us would agree that the overall quality of language learning can only be raised by a thorough acquaintance with the principles, practice, and theory of language testing and assessment.

The second assumption is a conclusion that follows this first one. While language testing specialists are firmly convinced that what language testing has learned is very much worth transmitting, the opportunities to do so are few and far in between (with the notable, happy exception, of seminars such as this one). Now, someone might claim that what, with some dozen introductory language testing handbooks available in English alone, teachers who haven't had the chance to learn about testing during teacher training can simply go to their friendly, local bookstore, order one, and read it. Such a view is tenderly optimistic to the point of being naive.

All of us could make a long list of those textbooks in other areas that we will be reading sometime next week, no doubt. Or, perhaps next month. Or, more likely, considering all the work we have to do, sometime next summer. That is, of course, unless we have to move house, change diapers, get remarried, or read and grade all of those student's essays. Friends who are classroom teachers are usually so tired when they get home Saturday afternoon, that the last thing they want to do is to start reading a textbook on testing.

There is this Peter Pan view of language teaching that says, usually without thinking, that all teachers will somehow acquire expertise and knowledge that in some cases, even their professors have not had time to gain. To take this stance is hypocritical and unfair. Worse yet, it is unrealistic.

My second assumption, therefore, is that when it comes to the teaching of language testing to teachers, whether to those already in the schools or to those still hoping to get there, much of it, if not almost all of it, will have to be through non-traditional approaches and channels. Like this seminar, for example. Or through special, intensively advertised workshops. Or as adjunct meetings to regular pedagogy conferences. Or through series of introductory articles in popular, teachers' journals. But we do need to remember one thing. This is that the large gain in the number of professional, specialists' publications on language testing is not matched by a similar gain in the number of classroom teachers who have had specific and thorough instruction in language testing. The problem is to explain why they should spend much of their time learning about language testing when so often, the educational authorities do not see fit to make this central subject a required one.

This brings us to the third of the assumptions which are important when we talk about the teaching of testing. It is an irony, or at least a cause for some self-directed humor, that language testing specialists have spent so little of our time at conferences talking

about the teaching of language testing. Yet from the very beginning of language testing as a field, it was apparent that if language testing and assessment in the schools and universities was to improve, it had to be with the active cooperation of classroom teachers. It is worth remembering here that Lado's Language Testing: The Construction and Use of Foreign Language Tests had a subtitle, A Teacher's Book. And, as far as I'm aware, almost all language testing books that have followed, have followed Lado's example and have been written primarily for, and directed to, the teacher.

It is also appropriate here to note that the Interuniversitäre Sprachtestgruppe, the IUS, was founded to pursue two goals. First, there was a need, at least in our educational venue, to establish an informal, frequently-meeting forum where specialists in language testing could talk to other specialists and thereby be aided and abetted in their efforts. Secondly, there was an obvious need to provide a friendly, informal setting in which non-specialists could acquire general to specific knowledge about language testing, and thereby be aided and abetted in their efforts, as well. We'd noticed that at most conferences - even those that had themes connected with teaching in the schools - there were few teachers and fewer students in attendance. Specialists talking to specialists, professors to professors, was the rule. We had also noticed that teachers who did not really know much about testing were understandably unwilling to stand up at formal meetings and ask questions that might be considered naive. Finally, we were well aware that the inherent interest in language testing would not be sufficient to have many paths beaten to our international doors. Therefore for all these reasons, the IUS would need to be a welcoming, rather than a closing or closed, organization.

What should be well-understood, especially as it is of fundamental importance to the teaching of testing, is that this second goal did not just spring from altruistic concerns or, God forbid, from any warped sense of public relations. Rather, the two goals are intertwined, at least in educational systems in which teachers have a say in what is taught and how what is taught is tested. No matter how good our tests are, no matter how refined, how attentive to current theory, psychometric principles, and pedagogical good house-keeping, they really don't stand much of a chance without the active understanding and support of the classroom teacher.

We're not just concerned with newly developed or better testing techniques and measures. As has often been said, and is still so obviously true, the worst testing practices that are so commonly encountered do not result from any intent on the part of those who use them to be unfair to pupils. Rather, they are generally traditional in that reliability is of little interest. And authority is usually more than sufficient to take care of any embarrassing questions about consistency or content. It is therefore all rather simple, it might seem. All one must do is to bring the two groups together, the testers and the teachers.

Or so it might seem. About three or four weeks ago, I just happened to be present during one of those discussions among col-

BEST COPY AVAILABLE

leagues that the outsider might see as a heated debate, or even as an argument. The topic of translation testing, that is, the grading and scoring of translation examinations, somehow came up that morning between classes and coffee. I was just an observer; if not neutral, at least silent, for I wanted to hear what was said about translation testing. For some time I'd been trying to find out more about this subject, still so important in our university system. I'd been checking with specialists in the area, people who had been trained both as translators and as testers. These specialists had been in general agreement that injecting reliability into translation grading would encounter the usual problems of what constitutes validity-based variance, and what error, and so on. But, following the usual analytic or holistic approaches to grading long used with essays, establishing a high degree of reliability posed no real problems, at least from the language tester's point of view. In short, the argument did have my full attention, as these colleagues were much more sure of themselves and their positions than the specialists I'd talked with had been.

What it all came down to was that those present that morning did not believe that testing had anything worthwhile to say about translation grading. One colleague stated in no uncertain terms that all such testing approaches he'd read about were essentially nonsense (although that was not the exact word he used). He had taught for years and years. He had read more translations than he cared to remember. He had graded translation examinations until his eyes glowed red, like some neon rabbit. And he knew for a fact that the grading of translation was subjective in fact and in practice, that no consistency could be reached, and that nothing, let alone a few correlational coefficients would ever change that. There was no use even talking about seeing new ways to score translations, and that was that, period,full-stop, punkt!

One factor we should keep in mind, therefore, is that language testing is frequently an emotional issue. A great many people already have very strong opinions on what is good and bad, possible and impossible. Moreover, those who are strongest in their opinions are often those who actually know the least about language testing.

As much as we would counter these negative feelings, we can't overlook the very very positive benefits of securing teacher support and understanding. These can equally be given example, and two come readily to mind. First is the incredibly wide-acceptance in North America in foreign-language teaching circles of what many of us know best as the FSI Oral Interview procedure. The second example is the graded objectives movement in Great Britain which we've had the pleasure of hearing more about during this seminar. What is striking about both movements, although an ocean apart, is not, as one might expect, the actual tests or testing approaches. Both have their antecedents in the established testing literature, so what is being done in either area is not really new. What is so striking is that each has spread and has been so seemingly successful because each has been actively supported by classroom teachers. This did not come about by chance.

Specialists working with the oral interview, for example, spent several years presenting this approach through just about any channel available, both professional and popular. Groups such as ACTFL, the Interagency Language Roundtable, TESOL, ETS, and individuals working in schools and universities came together to explain the oral interview and to answer questions. Teachers were gone to, talked with, and listened to when they questioned this problem or that detail. The results have been simply astounding.

We should note that this publicizing has not in any way been like the selling of a soap, a new deodorant, or the latest toothpaste. Those most responsible for presenting the oral interview have also been most responsible in pointing out its limitations, constraints and, above all, the great effort and training needed to make it effective. This nationwide movement in the US, therefore, is far from being a mindless surge. We need to mention this last factor, for there is always the specialist's fear that in making anything better known to the non-specialist, it will necessarily be simplified, or cheapened in some way.

Experienced language teachers, however, are already skeptical about any teaching approach that claims to be the one best way of teaching anything or anyone. Most of us, for instance, are already past counting on one hand the number of times a so-called new approach or method has been proclaimed as the quickest and surest way to language learning Nirvana. Most of us have also discovered from experience in the classroom what others have established through research: no single teaching method fits everyone teaching or everyone taught.

There are of course in language testing - as in any field - those who would sell a testing product that cures all academic illnesses in record time, that can be administered anytime by anyone for any reason to everyone. But language testing specialists are in a good position to compare the one to the other - the teaching cure-all to the testing cure-all - and thereby lessen the appeal of both.

This third assumption, therefore suggests that probably the single best way to improve the quality of testing in schools is not necessarily by introducing new tests or new testing techniques. Rather, it is to make available, to make better known, the critical tools already available and used in language testing.

I'd like to pause here for a moment, and sum up some of the points I've tried to make thus far in talking about the teaching of testing. First, language testing specialists are in agreement that their field - at its broadest - is at the heart of linguistic, language learning, and language teaching research and methodology. More specifically, they are convinced that the intelligent application of language testing approaches, techniques, and methods to the classroom can do a great deal to improve the overall quality of language learning and teaching. However, at the same time, this viewpoint is not shared for the most part where it most counts. A relatively small number of in-service teachers have been specifically trained in

language testing. A relatively small number of teacher training programs require specific coursework in language testing and assessment. The language teaching community might pay lip service to the importance of language testing and here and there offer active support. Nontheless, overall, few language testers have the opportunity to teach more than a little bit about testing to more than a few future teachers.

Secondly, there is already a large amount of literature about language testing set at the introductory level, and specifically directed at the modern language teacher working in a second, or foreign language classroom. For the most part, this literature is readily available. In some cases, free materials have been made available by various non-profit organizations. However, at the same time, because language testing is seldom given the same five-star ranking accorded other disciplines in teacher training, it cannot simply be assumed that the practicing teacher will see fit to purchase and read these materials. Certainly, other disciplines compete for his or her time. And certainly, teachers always have less time than might appear to those who do not teach.

Thirdly, and this should be stated as bluntly as possible, language testers cannot go it alone. We cannot assume that what we say and write will somehow naturally be of interest to those outside our specialty, just because we think it should be. Specialists might all be in agreement that a certain new test is damned fine, or another one potentially as dangerous as the Black Plague in its effects on pedagogy or the pupil. But this means very little in either the success of the one, or, sadly, the eradication of the other. This is no doubt a hard lesson for us to learn. But the fact that in educational systems marked by great overall sophistication and strong tradition, markedly unfair examinations still persist - and sometimes are even dominant - should forcibly remind us that cooperative action is not nice, but ultimately necessary. There are, as well, strong attitudes, both negative and positive towards types of testing and testing approaches already in existence. As a result, language testing must be prepared to go out and deal with what it considers to be unfounded prejudices, misinformation, or even, clichès and stereotypes about testing that are of the common garden variety. Every good test must have 100 questions, right? Of course not.

Basing these assumptions upon previous literature written on the teaching of testing, research into attitudes towards tests and testing, and conversations with language testing specialists internationally, this is the situation, generally speaking, that we must deal with when we teach, or would teach, about language testing. More or less, of course, for each educational system, each teacher training program, has its own tendencies and trends. A few do require specific training in testing (Israel is a good example). Others simply list it as one of those abilities teachers should have. Still others would not even list any language testing book on the standard "important or required" reading lists distributed to all incoming students. Macbeth, yes. Oller, no.

Someone once said that none of us [illegible] to take a good, long, realistic look at anything, be it the state of the contemporary world, our personal finances, our children's rooms, or even the teaching of testing. I do think it would be wrong, however, to be anything but guardedly optimistic, or even out and out positive, about what can be done, and in some systems, is being done in the teaching of testing despite this overall situation that I've briefly described. Again, you are a very good sign of the times, as it were, a most positive indication that language testing is being viewed more and more as a very worthwhile area for the classroom teacher to spend time in and effort on. Someone is buying those language testing volumes. No one really expects Hudson's new (1988) volume on testing, published by Oxford University Press, for example, to meet the best seller levels of a novel like Fear of Flying. But it has appeared, another one of many good ones, and will be read by many.

Also operating in favor of language testing are two general but continuing educational trends. First, and perhaps most important, there is that concept known as accountability. Simply stated, more pressure, whether coming from governments, tax payers, parents or pupils, is being placed on those who teach to be responsible and public about their successes and failures. In some systems, for instance, teachers are now required by law to demonstrate their own proficiency in the language they teach. In others, teachers are paid to the extent that they can demonstrate that their pupils have achieved what they should have achieved.

The second general trend affecting a stronger interest in testing has made itself more and more felt since the end of the Second World War. The spread of education at all levels to groups not previously encompassed has also led to a general breakdown of that authority that so characterized traditional testing approaches in the schools. Again, simply stated, fewer parents, or pupils, or students are willing to accept the teacher's word as being the final word of God and government.

Over twenty years ago now, an acquaintance of mine teaching at a university in Michigan was taken to court by a student who thought that she deserved a "B" on an essay examination she'd written, when he'd given her only a "D." Experts were called in to testify about reliability. The essay was read by groups of specialists to determine inter- and intra-rater reliability ranges. He won the court case, finally, but it didn't do much for his nerves. Now, you might be thinking, that sort of nonsense can't happen here, wherever "here" might be. But it did happen there, then, and had, as you can well imagine, a very strong effect on the standards and care of essay grading. Suddenly, testing was very relevant, and how one can grade essays reliably a topic of more than casual or intellectual interest. Students not only take tests; they take them personally.

These two general trends might be seen as constituting negative sources of motivation to learn more about testing. True, but I've met quite a few language testers who, having tried to explain the importance of reliability and such to sceptical audiences, feel that those who have ignored such questions so intimately related to ethics

and fairness deserve what they get. There's nothing personal, mind you, it's just that they feel that certain practices have continued too long to be tolerated any longer. If they're forced to choose between the teacher's comfort and the pupil's welfare, most testers feel they have no choice.

What various legal authorities, above all in North America, are now saying about classroom testing is a most interesting subject. But here, it just needs mentioning as one of the several factors that is creating more of an interest in the literature of language testing. The point is that despite the handicaps we must realistically face in the teaching of testing, much can be done, and much must be done.

In the time left, I would therefore like to turn to what we might call the content material and matter. Now, you are very sensitive, as the result of lectures during this seminar, about content validity. Aware of this, I will state at the outset, that the suggestions I have for various areas that should be more stressed in teaching testing do not claim any universal, or specific validity for any given educational system. Importantly, they are offered as fundamental bases and considerations in the teaching of testing, considerations that have been too often slighted in more usual approaches to the topic. I've chosen four areas that I think we should emphasize more than we have.

As content in the teaching of language testing, I do believe and, over time, have become more firmly convinced, that the concepts of reliability and validity should be first and foremost in any list of what must be understood. "Understood" is the key word here, for a good feel for the concepts and their interrelationship is, in the long run, much more important than a quick review and a few exercises in how one, for instance, calculates a test-retest statistic. Just about everything else in language testing does intersect at some point with these two, and time spent insuring that students do have a good feel for them is well spent, indeed. Both need to be given their full due as the central concepts and not just treated in passing, or placed as afterthoughts.

There is another reason for according reliability and validity this central and primary position. And not understanding this is in fact why the topic of reliability and validity is so often greeted with a "yeah, yeah, we've already heard of that" sort of reaction. What separates the traditional testing approaches from the modern, the language tester from the metrically-naive layman, is not specific types of tests or testing approaches. It is an appreciation of the nature of, and the need for, reliability. This is encompassed in that measurement dictum that "a test can be reliable without being valid, but to be valid, it must be reliable."

When this concept is grasped, the inter-relationship appreciated, a sort of "aha!" reaction often sets in, and the basic importance of language testing becomes clearer. It is an ice-breaker, if you wish, as a topic. It would not be too much of an exaggeration to say that teachers are converted, at this point, and never quite look at testing

in the same way again. One thing that Alan Davies was getting at in the first issue of Language Testing (1984:112) can be furthered here. It is very easy, especially with computers just about everywhere, to introduce the topic of reliability, and soon have people turning out data. But we should not confuse the ability to perform various statistical manipulations with an thorough understanding of the concepts. It certainly helps, but it's not the same.

To explain reliability so that students feel it, takes a good teacher and a good bag of metaphors and analogies. For example, I've noticed that in Harrison's most recent introductory language testing volume (1983:10f.), he uses an elastic measuring tape as a metaphor. "There would be little point in trying a measure people's waists with a piece of elastic," he notes. And, "the tape measure should be placed so that it goes around people's waists, not round their hips or the waistband of the trousers they are wearing."

Reliability and validity do have the reputation of being esoteric areas, and those who think that the contemplation of construct validity theory is good for the soul probably don't have many sympathizers. But the general concepts can be introduced in a relatively short period of time, and the effect of their being understood is a dominant one on whatever else follows.

A couple years ago, for instance, Ulli Maetz took a fresh group of students at a workshop in Bielefeld through the concept, and pointed out what reliability had to say about our grading system in German schools. The atmosphere was electric when the two were brought together. "That can't be true," was of course one reaction to be expected. The other one, sort of "well I'll be damned" (and quite a few pupils obviously have been) was another. But at this point, how one can express and determine reliability made more sense, for there then existed clear reasons for doing so. For validity, it is perhaps enough here to repeat Spolsky's well-known phrase (1975:153), "the central problem of foreign-language testing, as of all testing is validity."

If we take reliability and validity and their interrelationship together, then it can be stated that an understanding among teachers if their significance can do more good, more quickly, than anything else to improve the quality of testing in the schools. Everything else, all that other stuff, more or less follows these principles, with the possible exception, depending upon definitions, of the "why" of testing. It is most unfortunate, I think, that reliability and validity have come only to be associated among many in language areas with only certain types of testing, as if they were things that Lado and his followers thought up to make our lives more difficult. On the other hand, if we look at standard introductory testing volumes, it is as if language testers, having accepted these principles, assume somehow that everyone else has, as well. They are often presented as givens, instead of as concepts that must be continually presented, defended, and explained to each new generaltion of teachers. In any teaching of testing, they must have first place, always.

The second area of language testing that we need to emphasize more than we have goes in the opposite direction, that is, from the heart of the forest to the trees. This is what we could call the overall educational and social context, and functioning of testing. The type thing Spolsky did with his three trends in language testing article would fit here, of course. But what I'm thinking of is a larger historical perspective, the type thing more often done in educational and psychological measurement courses, where the field of human measurement is surveyed and exampled. If we wish, we could call this a perspective on the "why" of language testing, today and yesterday, in this place, and that.

We would want to include as much comparative information as is possible, that is, what systems other than ours are doing and why. Those lectures we've heard here about testing in different educational systems, for instance, are of great value to those of us who will, when all is said, not have to work within them. They offer us new ways of looking at our own cases. They challenge our own assumptions. Furthermore, and this is perhaps most important, they tend to knock down that stubborn view - the one I exampled earlier with reference to translation testing - that something simply can't be done, or isn't worth doing.

For years, we were told that oral interviews simply couldn't be carried out on a continent-wide basis and still have any comparative value. But this is what appears to be happening in the US today. In another direction, the work on self-assessment measures done by Tibor von Elek and his colleagues often comes as a pleasant surprise to those who think a testing device must always assume the examinee is somehow lying or, at best, untrustworthy. As a third example, there is the TOEFL. What it is and what it does just doesn't make sense to anyone who doesn't know the background of educational needs in North America that made it necessary. Too often it has been judged simply as a test in isolation, and not as part of a larger measurement procedure in a larger educational context.

Yet we should not just begin with our own time or just within our own systems. It's too easy to be parochial in both, to assume what we do is somehow natural, or the only way. A standard article given in introductory educational and psychological measurement courses is entitled "A test-dominated society: China 1115 B.C. - 1905 A.D." Americans, who might otherwise see their own nation-wide, testing programs as both recent and singular, usually benefit greatly by viewing a society that had similar programs some two-and-a-half thousand years before Columbus got lost on his first trip to India. We would also want students to see and, if possible, take as many different types of tests as possible. "Do onto yourself, before you do onto others," is a good testing rule. We would always want to survey what is included in the broad range of interests we call language testing. Just to show how much is available is important. It also tends to prevent us from going out and constructing what we think are new measures, before we've taken the time to see what is already available.

The third area that I would suggest needs to be given much more attention is related to this second one. This is what is commonly

termed language learning research. Here, we would want to point out that classroom language testing is only one small area where language tests strongly influence language teaching and learning. Almost every single research study we can think of has, somewhere, usually at its heart, a language test of some type that serves as the fulcrum upon which teaching recommendations will be based, and new approaches claimed and defended.

What appears in the standard, international language teaching journals is fertile field for such material. Many of our students, for example, now find themselves taking courses about second and foreign language acquisition. There is that famous, or to many language teachers, infamous Bilingual Syntax Measure that sits right in the middle of many interlanguage theories and research studies. It's validity is a central question. Attitudinal and motivational research is fast becoming of greater interest in many systems. We have Hauptschule students who are now required to take another year of English, for instance, and what might still interest them, is a good question indeed. Adult education, bilingualism, and those languages losing their fair share of students to English have also accelerated interest in this area. Again, right in the middle are all those tests and testing techniques. Are they reliable? Are they valid? Are they appropriate?

We could go on, certainly, but it should be clear that language testing is a most useful applied tool for anyone, whether or not he or she will ever actually construct a test battery. It's interesting to look back over recent issues of popular journals, those supposedly directed at teachers such as the TESOL Quarterly. More and more, the pages are filled with statistics and references to various measurement concepts. Not a few of our students, probably most, simply skip across those pages, those correlational tables and such, just as if they were written in a foreign language that they didn't know. Here language testing can perform a badly needed interpretive service. It is wrong that many teachers should be unable to understand, let alone judge, so much of what is presented today. The misuse of testing, tests, and measurement concepts is a very big problem and here we can do much, rather easily, to restrict it, as well.

There is one other aspect to this emphasis that is less applied, but just as important, I think. This is the issue of complexity. An increasing methodological rigor provided by testing in all areas of language studies has led to an increasing awareness of just how complex it all is. Ironically, if someone is looking for a simple, straight-forward answer - and who isn't? - there is a natural tendency to shy away from this complexity. Often, as a result, both research and tests which promise the most are also those which are least sophisticated in measurement terms. They are, therefore, the most appealing to those with not much background in testing. We need to deal with this problem head on, and why, for instance, a test that is not pretested, or does not report say, reliability statistics or even SEMs is probably not a good one, is usually a good place to start.

This brings us to the fourth and final of those areas that I

think we need to stress more within that "all of it" possible universe of content. It gets to the heart of why so many of us are so persistent in talking about the teaching of testing. It brings together, as well, our concern with the quality of testing and our emphasis on reliability and validity. This is the entire question sets of ethics and accountability and fairness.

All language testers must operate, as they know, within a set of what might be called functional ideals. There is, for example, no such thing as a perfectly reliable test. Knowing this, we still manipulate the concept in theory and practice as if there were. Likewise, there is no such thing as a perfectly valid test for all purposes, pupils and conditions. Yet we also manipulate this ultimate constraint as if there could be. We do the best we can, knowing all the time that from the very beginning, the nature of testing and the extent of our knowledge about what is to be tested will compromise it. We know from the outset that it will be far less than ideal.

One reason language testers have always been seriously concerned with the teaching of testing is that perhaps more than others, they know what a tricky business it all is, even at its best. With a constant eye on the educational and social consequences of testing, they must, as a matter of professional and personal conscience, continually point out the distance between best practice and ideal. A good knowledge of reliability and validity, therefore, also imposes upon one an awareness of the responsibility that connects the why, what and how of testing with decisions based upon it. And certainly, knowing the iffiness of the entire structure we are much much better off than if we didn't. We might not be much happier, or more content, but we are more responsible for what we teach and, more importantly, to those we teach.

The direct connection between what are often disparagingly termed "only" psychometric considerations, and the effects testing has on what is taught, who teaches, and who is taught, is also one reason that language testing is a rather exciting place to be. By its very nature, it tends to challenge tradition, authority, and not least, simple smugness. Those coming to testing for the first time often expect a somewhat reactionary activity, and are therefore often surprised just what revolutionary potential is always there. Approaching the field of language testing through discussions of fairness, and accountability, is often not a bad way to start, for it sparks off a greater interest in what it's all about.

To conclude here, we should state of course, that everything else in that "all of it" of language testing is also important and, given time and world enough, we would also teach. However, the four areas which I've suggested we might want to give more emphasis to are seen as basic structures upon which the rest can be built. Even if a classroom teacher is never called upon to construct a test battery, even if a teacher never carries out his or her own empirical research, they are highly useful. Overall, I hope I've been able to strengthen our mutual view that the qualitiy of testing stands to gain as much from the teaching of testing as it does simply through more testing.

And finally, that "reliability, and validity, and you know, all that stuff," is of central importance to any teacher, any where, any time.

Notes

1 This was the summary paper for the Hoinola teachers' meeting, and the oral-presentation style has been preserved here.

Bibliography

Davies, A. 1984. Review of Oller, J.W., Jr. Issues in language testing research. Language Testing 1/1: 111-115.

Harrison, A. 1983. A language testing handbook. London: Macmillan Press.

Madsen, H. 1983. Techniques in testing. New York: Oxford University Press.

Spolsky, B. 1975. "Language testing - the problem of validation." In Palmer, L. and Spolsky, B. (eds). 1975. Papers on language testing 1967-1974. Washington, D.C.: Teachers of English to Speakers of Other Languages. 146-153.

Stevenson, D.K. 1983. "Foreign language testing: all of the above." In James, C.J. (ed). 1983. Practical applications of research in foreign language teaching. Lincolnwood, IL: National Textbook Company. 153-203.